

共起情報を考慮した $tf \cdot idf$ 法に基づく 関連文書間の自動ハイパーテキスト化

岡村 潤 大森 信行 山口 登志実 森 辰則 中川 裕志

横浜国立大学 工学部 電子情報工学科

E-Mail: {jun,ohmori,yamaguti,mori}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

概要

今日、高機能・高性能なシステムに付属されるマニュアルは別冊に分割されていることが多く、利用者が必要とする情報を効率よく得ることは困難となっている。本論文では、複数マニュアル間の関連箇所を自動的に抽出しハイパーリンクを生成する方法について述べる。本研究では、マニュアル内の段落を1つの文章のまとまりと考え、段落中のキーワードに $tf \cdot idf$ 法によって重みをつけ、ベクトル空間法により段落間の類似度の大きさを求め、その関連度にしたがって段落間でのハイパーリンクを自動生成する。また、より高精度な関連付けのために段落間でのキーワードの共起情報とキーワードの語彙連鎖を用いることが有効であることが確かめられた。

1 はじめに

今日、各種機器やソフトウェアが高機能・高性能になるにつれて、それに付属されるマニュアルも利用者のレベルや使用用途別に合わせて、複数に分冊されるような形態をとるようになってきた。これに伴いユーザもそれらのマニュアル群からユーザ自身が欲する知識・概念を取り出すことが必要になってきている。そのような必要な知識・概念を得るのに、従来の紙面のマニュアルにおいては参照すべき部分を目次や索引から探しだし、読み進めていくことによってなされていた。しかしながら複数に分かれたマニュアル群において、目次や索引から参照箇所を探し出す作業はユーザに大きな負担をかけることになる。

このような負担を軽減するために、最近では文書や図の間の相互参照情報(ハイパーリンク)をもつハイパーテキスト型マニュアルが見受けられるようになってきた。しかし、このハイパーリンクの構築はあらかじめ人間の手作業によって行なわれる必要があり、大規模マニュアル群におけるこの作業は困難を極める。本稿では、複数マニュアル間におけるハイパーリンクを情報検索手法を用いて自動的に生成するシステムを提案する。マニュアルにおいては語句の説明箇所の他に、一連の操作手順などが書かれたまとまった文書が参照対象になり得る。例えば、初心者用マニュアルに例示されている操作について、それに対応する詳細記述をリファレンスマニュアルで調べる場合などが想定される。そこで本稿では、図1に示すような節や項などのまとまった文書単位での自動ハイパーリンク生成を考える。

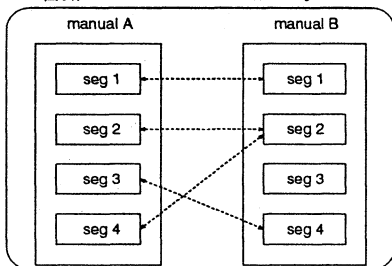


図1: システムが生成するハイパーテキストの概念図

2 自動ハイパーテキスト生成システム

本システムにおいて、我々は自動ハイパーテキスト生成を次のように実現する。

1. ハイパーリンク生成の対象は、文書小単位(セグメント)である。2つのマニュアルをセグメント単位に区切り、その全てを候補と考える。セグメントの単位としては、意味的なまとまりを考慮し、節、項とする。
2. 関連付けについては、まず両マニュアルにおける任意の組合せについてあらかじめ類似度計算を行っておく。ハイパーリンクは利用者に提示する時に、類似度の高いものから動的に生成し、提示する。

1. については、HTML など構造をもつ記述形式になっていれば、文書構造からセグメントを認識できるため容易に自動化できる。

2. については、類似度のスコア付けが問題となる。この類似度のスコア付けには、情報検索で広く用いられている、 $tf \cdot idf$ 法に基づくベクトル空間モデルを利用する。

2.1 システム概要

本システムは、図2に示す4つのサブシステムより構成されている。

本システムでは、以下のような手法でセグメント間類似度を計算する。

1. 対象となる文章群から語を抽出し、それぞれに $tf \cdot idf$ 値によって重要度を与える。
2. ベクトル空間モデルによってセグメント間の類似度を計算する。類似度計算においては、以下のようなベクトルを生成する。
 - 一つのセグメントに一つのベクトルを対応させる。
 - ベクトルの各次元には各単語が、各成分には対応する単語の重要度、すなわち単語の $tf \cdot idf$ 値が割り当てられる。

共起情報を考慮した $tf \cdot idf$ 法に基づく 関連文書間の自動ハイパーテキスト化

岡村 潤 大森 信行 山口 登志実 森 辰則 中川 裕志

横浜国立大学 工学部 電子情報工学科

E-Mail: {jun,ohmori,yamaguti,mori}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

概要

今日、高機能・高性能なシステムに付属されるマニュアルは別冊に分割されていることが多く、利用者が必要とする情報を効率よく得ることは困難となっている。本論文では、複数マニュアル間の関連箇所を自動的に抽出しハイパーリンクを生成する方法について述べる。本研究では、マニュアル内の段落を1つの文章のまとまりと考え、段落中のキーワードに $tf \cdot idf$ 法によって重みをつけ、ベクトル空間法により段落間の類似度の大きさを求め、その関連度にしたがって段落間でのハイパーリンクを自動生成する。また、より高精度な関連付けのために段落間でのキーワードの共起情報とキーワードの語彙連鎖を用いることが有効であることが確かめられた。

1 はじめに

今日、各種機器やソフトウェアが高機能・高性能になるにつれて、それに付属されるマニュアルも利用者のレベルや使用用途別に合わせて、複数に分冊されるような形態をとるようになってきた。これに伴いユーザもそれらのマニュアル群からユーザ自身が欲する知識・概念を取り出すことが必要になってきている。そのような必要な知識・概念を得るのに、従来の紙面のマニュアルにおいては参照すべき部分を目次や索引から探しだし、読み進めていくことによってなされていた。しかしながら複数に分かれたマニュアル群において、目次や索引から参照箇所を探し出す作業はユーザに大きな負担をかけることになる。

このような負担を軽減するために、最近では文書や図の間の相互参照情報(ハイパーリンク)をもつハイパーテキスト型マニュアルが見受けられるようになってきた。しかし、このハイパーリンクの構築はあらかじめ人間の手作業によって行なわれる必要があり、大規模マニュアル群におけるこの作業は困難を極める。本稿では、複数マニュアル間におけるハイパーリンクを情報検索手法を用いて自動的に生成するシステムを提案する。マニュアルにおいては語句の説明箇所の他に、一連の操作手順などが書かれたまとまった文書が参照対象になり得る。例えば、初心者用マニュアルに例示されている操作について、それに対応する詳細記述をリファレンスマニュアルで調べる場合などが想定される。そこで本稿では、図1に示すような節や項などのまとまった文書単位での自動ハイパーリンク生成を考える。

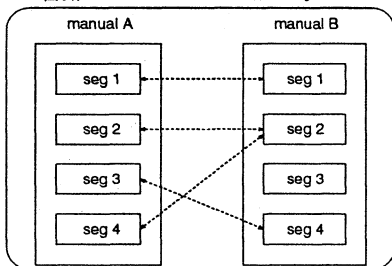


図1: システムが生成するハイパーテキストの概念図

2 自動ハイパーテキスト生成システム

本システムにおいて、我々は自動ハイパーテキスト生成を次のように実現する。

1. ハイパーリンク生成の対象は、文書小単位(セグメント)である。2つのマニュアルをセグメント単位に区切り、その全てを候補と考える。セグメントの単位としては、意味的なまとまりを考慮し、節、項とする。
2. 関連付けについては、まず両マニュアルにおける任意の組合せについてあらかじめ類似度計算を行っておく。ハイパーリンクは利用者に提示する時に、類似度の高いものから動的に生成し、提示する。

1. については、HTMLなど構造をもつ記述形式になっていれば、文書構造からセグメントを認識できるため容易に自動化できる。

2. については、類似度のスコア付けが問題となる。この類似度のスコア付けには、情報検索で広く用いられている、 $tf \cdot idf$ 法に基づくベクトル空間モデルを利用する。

2.1 システム概要

本システムは、図2に示す4つのサブシステムより構成されている。

本システムでは、以下のような手法でセグメント間類似度を計算する。

1. 対象となる文章群から語を抽出し、それぞれに $tf \cdot idf$ 値によって重要度を与える。
2. ベクトル空間モデルによってセグメント間の類似度を計算する。類似度計算においては、以下のようなベクトルを生成する。
 - 一つのセグメントに一つのベクトルを対応させる。
 - ベクトルの各次元には各単語が、各成分には対応する単語の重要度、すなわち単語の $tf \cdot idf$ 値が割り当てられる。

共起情報を考慮した $tf \cdot idf$ 法に基づく 関連文書間の自動ハイパーテキスト化

岡村 潤 大森 信行 山口 登志実 森 辰則 中川 裕志

横浜国立大学 工学部 電子情報工学科

E-Mail: {jun,ohmori,yamaguti,mori}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

概要

今日、高機能・高性能なシステムに付属されるマニュアルは別冊に分割されていることが多く、利用者が必要とする情報を効率よく得ることは困難となっている。本論文では、複数マニュアル間の関連箇所を自動的に抽出しハイパーリンクを生成する方法について述べる。本研究では、マニュアル内の段落を1つの文章のまとまりと考え、段落中のキーワードに $tf \cdot idf$ 法によって重みをつけ、ベクトル空間法により段落間の類似度の大きさを求め、その関連度にしたがって段落間でのハイパーリンクを自動生成する。また、より高精度な関連付けのために段落間でのキーワードの共起情報とキーワードの語彙連鎖を用いることが有効であることが確かめられた。

1 はじめに

今日、各種機器やソフトウェアが高機能・高性能になるにつれて、それに付属されるマニュアルも利用者のレベルや使用用途別に合わせて、複数に分冊されるような形態をとるようになってきた。これに伴いユーザもそれらのマニュアル群からユーザ自身が欲する知識・概念を取り出すことが必要になってきている。そのような必要な知識・概念を得るのに、従来の紙面のマニュアルにおいては参照すべき部分を目次や索引から探しだし、読み進めていくことによってなされていた。しかしながら複数に分かれたマニュアル群において、目次や索引から参照箇所を探し出す作業はユーザに大きな負担をかけることになる。

このような負担を軽減するために、最近では文書や図の間の相互参照情報(ハイパーリンク)をもつハイパーテキスト型マニュアルが見受けられるようになってきた。しかし、このハイパーリンクの構築はあらかじめ人間の手作業によって行なわれる必要があり、大規模マニュアル群におけるこの作業は困難を極める。本稿では、複数マニュアル間におけるハイパーリンクを情報検索手法を用いて自動的に生成するシステムを提案する。マニュアルにおいては語句の説明箇所の他に、一連の操作手順などが書かれたまとまった文書が参照対象になり得る。例えば、初心者用マニュアルに例示されている操作について、それに対応する詳細記述をリファレンスマニュアルで調べる場合などが想定される。そこで本稿では、図1に示すような節や項などのまとまった文書単位での自動ハイパーリンク生成を考える。

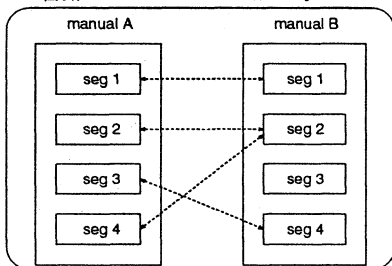


図1: システムが生成するハイパーテキストの概念図

2 自動ハイパーテキスト生成システム

本システムにおいて、我々は自動ハイパーテキスト生成を次のように実現する。

1. ハイパーリンク生成の対象は、文書小単位(セグメント)である。2つのマニュアルをセグメント単位に区切り、その全てを候補と考える。セグメントの単位としては、意味的なまとまりを考慮し、節、項とする。
2. 関連付けについては、まず両マニュアルにおける任意の組合せについてあらかじめ類似度計算を行っておく。ハイパーリンクは利用者に提示する時に、類似度の高いものから動的に生成し、提示する。

1. については、HTML など構造をもつ記述形式になっていれば、文書構造からセグメントを認識できるため容易に自動化できる。

2. については、類似度のスコア付けが問題となる。この類似度のスコア付けには、情報検索で広く用いられている、 $tf \cdot idf$ 法に基づくベクトル空間モデルを利用する。

2.1 システム概要

本システムは、図2に示す4つのサブシステムより構成されている。

本システムでは、以下のような手法でセグメント間類似度を計算する。

1. 対象となる文章群から語を抽出し、それぞれに $tf \cdot idf$ 値によって重要度を与える。
2. ベクトル空間モデルによってセグメント間の類似度を計算する。類似度計算においては、以下のようなベクトルを生成する。
 - 一つのセグメントに一つのベクトルを対応させる。
 - ベクトルの各次元には各単語が、各成分には対応する単語の重要度、すなわち単語の $tf \cdot idf$ 値が割り当てられる。

共起情報を考慮した $tf \cdot idf$ 法に基づく 関連文書間の自動ハイパーテキスト化

岡村 潤 大森 信行 山口 登志実 森 辰則 中川 裕志

横浜国立大学 工学部 電子情報工学科

E-Mail: {jun,ohmori,yamaguti,mori}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

概要

今日、高機能・高性能なシステムに付属されるマニュアルは別冊に分割されていることが多く、利用者が必要とする情報を効率よく得ることは困難となっている。本論文では、複数マニュアル間の関連箇所を自動的に抽出しハイパーリンクを生成する方法について述べる。本研究では、マニュアル内の段落を1つの文章のまとまりと考え、段落中のキーワードに $tf \cdot idf$ 法によって重みをつけ、ベクトル空間法により段落間の類似度の大きさを求め、その関連度にしたがって段落間でのハイパーリンクを自動生成する。また、より高精度な関連付けのために段落間でのキーワードの共起情報とキーワードの語彙連鎖を用いることが有効であることが確かめられた。

1 はじめに

今日、各種機器やソフトウェアが高機能・高性能になるにつれて、それに付属されるマニュアルも利用者のレベルや使用用途別に合わせて、複数に分冊されるような形態をとるようになってきた。これに伴いユーザもそれらのマニュアル群からユーザ自身が欲する知識・概念を取り出すことが必要になってきている。そのような必要な知識・概念を得るのに、従来の紙面のマニュアルにおいては参照すべき部分を目次や索引から探しだし、読み進めていくことによってなされていた。しかしながら複数に分かれたマニュアル群において、目次や索引から参照箇所を探し出す作業はユーザに大きな負担をかけることになる。

このような負担を軽減するために、最近では文書や図の間の相互参照情報(ハイパーリンク)をもつハイパーテキスト型マニュアルが見受けられるようになってきた。しかし、このハイパーリンクの構築はあらかじめ人間の手作業によって行なわれる必要があり、大規模マニュアル群におけるこの作業は困難を極める。本稿では、複数マニュアル間におけるハイパーリンクを情報検索手法を用いて自動的に生成するシステムを提案する。マニュアルにおいては語句の説明箇所の他に、一連の操作手順などが書かれたまとまった文書が参照対象になり得る。例えば、初心者用マニュアルに例示されている操作について、それに対応する詳細記述をリファレンスマニュアルで調べる場合などが想定される。そこで本稿では、図1に示すような節や項などのまとまった文書単位での自動ハイパーリンク生成を考える。

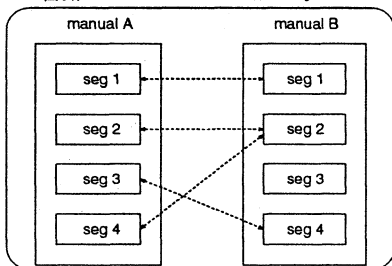


図1: システムが生成するハイパーテキストの概念図

2 自動ハイパーテキスト生成システム

本システムにおいて、我々は自動ハイパーテキスト生成を次のように実現する。

1. ハイパーリンク生成の対象は、文書小単位(セグメント)である。2つのマニュアルをセグメント単位に区切り、その全てを候補と考える。セグメントの単位としては、意味的なまとまりを考慮し、節、項とする。
2. 関連付けについては、まず両マニュアルにおける任意の組合せについてあらかじめ類似度計算を行っておく。ハイパーリンクは利用者に提示する時に、類似度の高いものから動的に生成し、提示する。

1. については、HTML など構造をもつ記述形式になっていれば、文書構造からセグメントを認識できるため容易に自動化できる。

2. については、類似度のスコア付けが問題となる。この類似度のスコア付けには、情報検索で広く用いられている、 $tf \cdot idf$ 法に基づくベクトル空間モデルを利用する。

2.1 システム概要

本システムは、図2に示す4つのサブシステムより構成されている。

本システムでは、以下のような手法でセグメント間類似度を計算する。

1. 対象となる文章群から語を抽出し、それぞれに $tf \cdot idf$ 値によって重要度を与える。
2. ベクトル空間モデルによってセグメント間の類似度を計算する。類似度計算においては、以下のようなベクトルを生成する。
 - 一つのセグメントに一つのベクトルを対応させる。
 - ベクトルの各次元には各単語が、各成分には対応する単語の重要度、すなわち単語の $tf \cdot idf$ 値が割り当てられる。